ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 1, January 2016

RDIB Technique for Degraded Document Images

Umesh Sonawane, Ajay Wagh, Bhandari Pritam, Harshada Pune, Pratik Kalamkar

B.E. Student, Dept. of IT, Raisoni College of Engineering, Ahmednagar, Maharastra, India

B.E. Student, Dept. of IT, Raisoni College of Engineering, Ahmednagar, Maharastra, India

B.E. Student, Dept. of IT, Raisoni College of Engineering, Ahmednagar, Maharastra, India

B.E. Student, Dept. of IT, Raisoni College of Engineering, Ahmednagar, Maharastra, India

Asst. Professor, Dept. of IT, Raisoni College of Engineering, Ahmednagar, Maharastra, India

ABSTRACT: Segmentation of text from badly degraded document images is a very challenging task due to the high inter / intravariation between the document background and the foreground text of different document images. In this paper, we using adaptive image contrast for define a novel document image binarization technique that addresses these issues. The adaptive image contrast is a combination of the local image contrast and the local image gradient that is tolerant to text and background variation caused by different types of document degradations. In the proposed technique, an adaptive contrast map is first constructed for an input degraded document image. To identify the text stroke edge pixels, contrast map is then binarized and combined with Canny's edge map. The document text is further segmented by a local threshold that is estimated based on the intensities of detected text stroke edge pixels within a local window. The proposed method is simple, robust, and involves minimum parameter tuning. It has been tested on three public datasets that are used in the recent document image binarization contest (DIBCO) 2009 & 2011 and handwritten-DIBCO 2010 and achieves accuracies of 93.5%, 87.8%, and 92.03%, respectively that are significantly higher than or close to that of the best performing methods reported in the three contests. Experiments on the Bickley diary dataset that define that compared with other techniques its challenging bad quality document images also show the superior performance of our proposed method.

KEYWORDS: Adaptive image contrast, document analysis, document image processing, degraded document image binarization, pixel classification..

I. INTRODUCTION

In the preprocessing stage of document analysis Document Image Binarization is performed well and it aims to differ the foreground text from background text.
For the ensuing document image processing tasks such as optical character recognition (OCR) a fast and accurate document image binarization technique is important. Though document image binarization has been studied for many years, the thresholding of degraded document images is still not resolved problem due to the high inter/intravariation between the text stroke and the document background across different document images. As illustrated in Fig. 1, the handwritten text within the degraded documents often shows a certain amount of variation in terms of the stroke width, stroke brightness, stroke connection, and document background. In addition, historical documents are often degraded by the bleed through as illustrated in Fig. 1(a) and (c) where the ink of the other side seeps through to the front. Imaging artifacts as illustrated in Fig. 1(e). These different types of document degradations tend to induce the document thresholding error and make degraded document image binarization a big challenge to most state-of-the-arttechniques.

The recent Document Image Binarization Contest(DIBCO) [1], [2] held under the framework of the International Conference on Document Analysis and Recognition (ICDAR) 2009 & 2011 and the Handwritten Document Image Binarization Contest (H-DIBCO) [3] held under the framework of the International Conference on Frontiers in

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 1, January 2016

Handwritten Recognition show recent efforts on this issue. We participated in the DIBCO 2009 and our background estimation method [4] performs the best among entries of 43 algorithms submitted from 35 internationalresearch groups. We also participated in the H-DIBCO 2010 and our local maximum-minimum method [5] was one of the top two winners among 17 submitted algorithms. In thelatest DIBCO 2011, our proposed method achieved second best results among 18 submitted algorithms. This paper presents a document binarization technique that extends our previous local maximum-minimum method [5] and the method used in the latest DIBCO 2011. The proposed method is simple, robust and capable of handling different types of degraded document images with minimum parameter tuning. It makes use of the adaptive image contrast that combines the local image contrast and the local image gradient adaptively and therefore is tolerant to the text and background variation caused by different types of document degradations. In particular, the proposed technique addresses.

The over-normalization problem of the local maximum minimum algorithm [5]. At the same time, the parameters used in the algorithm can be adaptively estimated

Fig1. Sample Degraded Images taken from DIBCO Series of images

II. LITERATURE SURVEY

Many thresholding techniques have been reported for document image binarization. As many degraded documents do not have a clear bimodal pattern, global thresholding is usually not a suitable approach for the degraded document binarization. Adaptive thresholding which estimates a local threshold for each document image pixel, is often a better approach to deal with different variations within degraded document images. For example, the early window-based adaptive thresholding techniques estimate the local threshold by using the mean and the standard variation of image pixels within a local neighborhood window. The main drawback of these window-based thresholding techniques is that the thresholding performance depends heavily on the window size and hence the character stroke width. Other approaches have also been reported, including background subtraction texture analysis, recursive method decomposition method, contour completion, Markov Random Field, matched wavelet cross section sequence graph analysis, self-learning, Laplacian energy user assistance and combination of binarization techniques. These methods combine different types of image information and domain knowledge and are often complex. The local image contrast and the local image gradient areVery useful features for segmenting the text from the document background because

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 1, January 2016

the document text usually has certain image contrast to the neighboring document background. They are very effective and have been used in many document image binarization techniques [5], [6], [7], [8]. In [14], the local contrast is defined as follows:

$$C(i, j) = Imax(i, j) - Imin(i, j)$$

where C(i, j) denotes the contrast of an image pixel (i, j), Imax(i, j) and Imin(i, j) denote the maximum and minimum intensities within a local neighborhood windows of (i, j), respectively. If the local contrast C (i, j) is smaller than a threshold, the pixel is set as background directly. Otherwise it will be classified into text or background by comparing with the mean of Imax (i, j) and I min(i, j). Bernsen'smethod is simple, but cannot work properly on degraded document images with a complex document background. We have earlier proposed a novel document image binarization method [5] by using the local image contrast that is evaluated as follows:
Where is a positive but infinitely small number that is added in case the local maximum is equal to 0. Compared with Bernsen's contrast in Equation 1, the local image contrast in Equation 2 introduces a normalization factor (the denominator) to compensate the image variation within the document background. Take the text within shaded document areas such as that in the sample document image in Fig. 1(b) as an example. The small image contrast around the text stroke edges in Equation 1 (resulting from the shading) will be compensated by a small normalization factor (due to the dark document background) as defined in Equation 2.

III. PROPOSED SYSTEM

In This section describes the proposed document image binarization techniques. Given a degraded document image, an adaptive contrast map is first constructed and the text stroke edges are then detected through the combination of the binarized adaptive contrast map and the canny edge map. The text is then segmented based on the local threshold that is estimated from the detected text stroke edge pixels. Some post-processing is further applied to improve the document binarization quality.

A. Contrast Image Construction:

The image gradient has been widely used for edge detection and it can be used to detect the text stroke edges of the document images effectively that have a uniform document background. On the other hand, it often detects many on stroke edges from the background of degraded document that often contains certain image variations due to noise, uneven lighting, bleed-through, etc. To extract only the stroke edges properly, the image gradient needs to be normalized to compensate the image variation within the document background. In our earlier method [5], the local contrast evaluated by the local image maximum and minimum is used to suppress the background variation as described in Equation 2. In particular, the numerator (i.e. the difference between the local maximum and the local minimum) captures the local image difference that is similar to the traditional image gradient The denominator is a normalization factor that suppresses the image variation within the document background. For image pixels within bright regions, it will produce a large normalization factor to neutralize the numerator and accordingly result in a relatively low image contrast. For the image pixels within dark regions, it will produce a small denominator and accordingly result in a relatively high image contrast. However, the image contrast in Equation 2 has one typical limitation that it may not handle document images with the bright text properly. This is because a weak contrast will be calculated for stroke edges of the bright text where the denominator in Equation 2 will be large but the numerator will be small. To overcome this over-normalization problem, we combine the local image contrast with the local image gradient and derive an adaptive local image contrast as follows:

$$Ca(i, j) = \alpha C(i, j) + (1 - \alpha)(Imax(i, j) - Imin(i, j)) \quad (3)$$

where C(i, j) denotes the local contrast in Equation 2 and (Imax(i, j) − Imin(i, j)) refers to the local image gradient that is normalized to [0, 1]. The local windows size is set to 3 empirically. α is the weight between local contrast and local gradient that is controlled based on the document image statistical information. Ideally, the image contrast will be assigned with a high weight (i.e. large α) when the document image has significant intensity variation. So that the proposed binarization technique depends more on the local image contrast that can capture the intensity variation well and hence produce good results. Otherwise, the local image gradient will be assigned with a high weight. WhereStd denotes the document image intensity standard deviation, and γ is a pre-defined parameter. The power function has a

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 1, January 2016

nice property in that it monotonically and smoothly increases from 0 to 1 and its shape can be easily controlled by different γ .γ can be selected from [0,∞], where the power function becomes a linear function when γ = 1. Therefore, the local image gradient will play the major role in Equation 3 when γ is large and the local image contrast will play the major role when γ is small. The setting of parameter γ will be discussed in Section IV. Fig. 2 shows the contrast map of the sample document images in Fig. 1 (b) and (d) that are created by using local image gradient, local image contrast [5] and our proposed method in Equation 3, respectively.For the sample document with a complex document background in Fig. 1(b), the use of the local image contras Produces a better result as shown in Fig. 2(b) compared with

Fig. 2. Contrast Images constructed using (a) local image gradient

variation within the text strokes, the use of the local image contrast removes many light text strokes improperly in the contrast map as shown in Fig. 2(b) whereas the use of local image gradient is capable of preserving those light text strokes as shown in Fig. 2(a). As a comparison, the adaptive combination of the local image contrast and the local image gradient in Equation 3 can produce proper contrast maps for document images with different types of degradation as shown in Fig. 2(c). In particular, the local image contrast in Equation 3 gets a high weight for the document image in Fig. 1(a) with high intensity variation within the document background whereas the local image gradient gets a high weight for the document image in Fig. 1(b). The result by the local image gradient as shown in Fig. 2(a) (Because the normalization factors in Equation 2 helps to Suppress the noise at the upper left area of Fig. 2(a)). But For the sample document in Fig. 1(d) that has small intensity Variation within the document background but large intensity.

B. Text Stroke Edge Pixel Detection:

The purpose of the contrast image construction is to detect the stroke edge pixels of the document text properly. The constructed contrast image has a clear bi-modal pattern [5], where the adaptive image contrast computed at text stroke edges is obviously larger than that computed within the document background. We therefore detect the text stroke edge pixel candidate by using Otsu's global thresholding method. For the contrast images in Fig. 2(c), Fig. 3(a) shows a binary map by Otsu's algorithm that extracts the stroke edge pixels properly.As the local image contrast and the local image gradient are evaluated by the difference between the maximum and minimum intensity in a local window, the pixels at both sides of the text stroke will be selected as the high contrast pixels. The binary map can be further improved through the combination with the edges by Canny's edge detector, because Canny's edge detector has a good localization property that it can mark the edges close to real edge locations in the detecting image. In addition, canny edge detector uses two adaptive thresholds and is more tolerant to different imaging artifacts such as shading. It should be noted that Canny's edge detector by itself often extracts a large amount of non-stroke edges as illustrated in Fig. 3(b) without tuning the parameter manually.In the combined map, we keep only pixels that appear within both the high contrast image pixel map and canny edge map. The combination helps to extract the text stroke edge pixels accurately as shown in Fig. 3(c).

C. Post-Processing:

Once the initial binarization result is derived from Equation 5 as described in previous subsections, the binarization result can be further improved by incorporating certain domain Knowledge First, the isolated foreground

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 1, January 2016

pixels that do not connect with other foreground pixels are filtered out to make the edge pixel set precisely. Second, the neighborhood pixel pair that lies on symmetric sides of a text stroke edge pixel should belong to different classes (i.e., either the document background or the foreground text). One pixel of the pixel pair is therefore labeled to the other category if both of the two pixels belong to the same class. Finally, some single-pixel artifacts along the text stroke boundaries are filtered out by using several logical operators as described in [4].

IV. EXPERIMENTAL RESULT

We have used our system on various on various type of images, like novel, books, and records, historical literature. Some of the results are stated as follow:

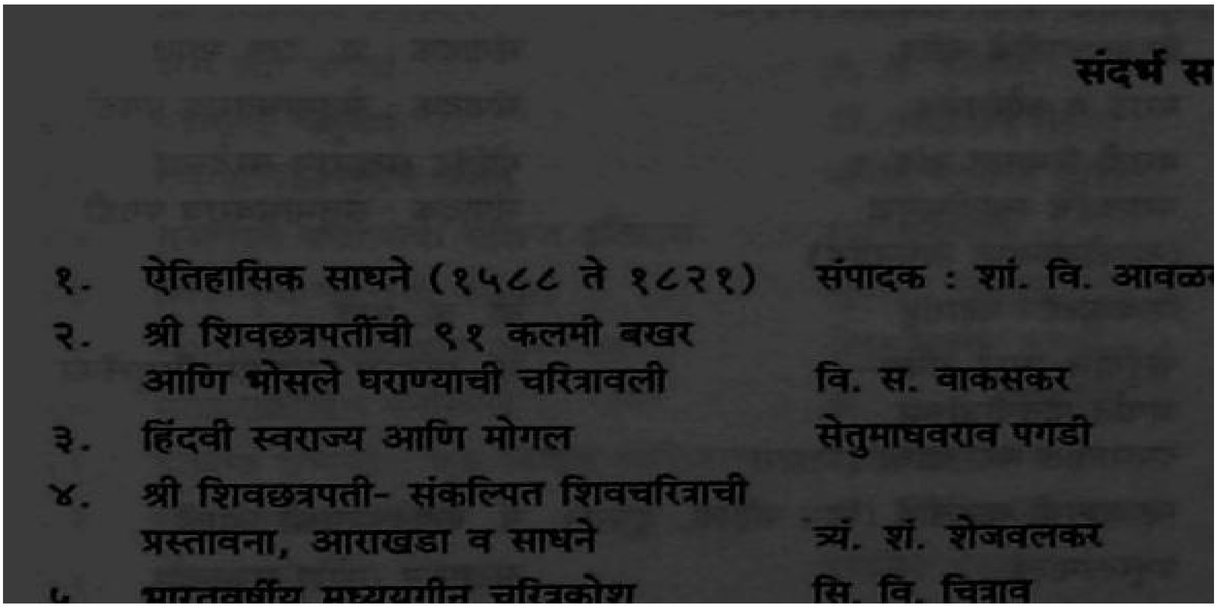
संदर्भ स
१. ऐतिहासिक साधने (१५८८ ते १८२१) संपादक : शां. वि. आवळ
२. श्री शिवछत्रपतींची ९१ कलमी बखर
आणि भोसले घराण्याची चरित्रावली वि. स. वाकसकर
३. हिंदवी स्वराज्य आणि मोगल सेतुमाधवराव पगडी
४. श्री शिवछत्रपती- संकलित शिवचरित्राची
प्रस्तावना, आराखडा व साधने त्र्यं. शं. शेजवलकर

Fig. 3: Input image for proposed system

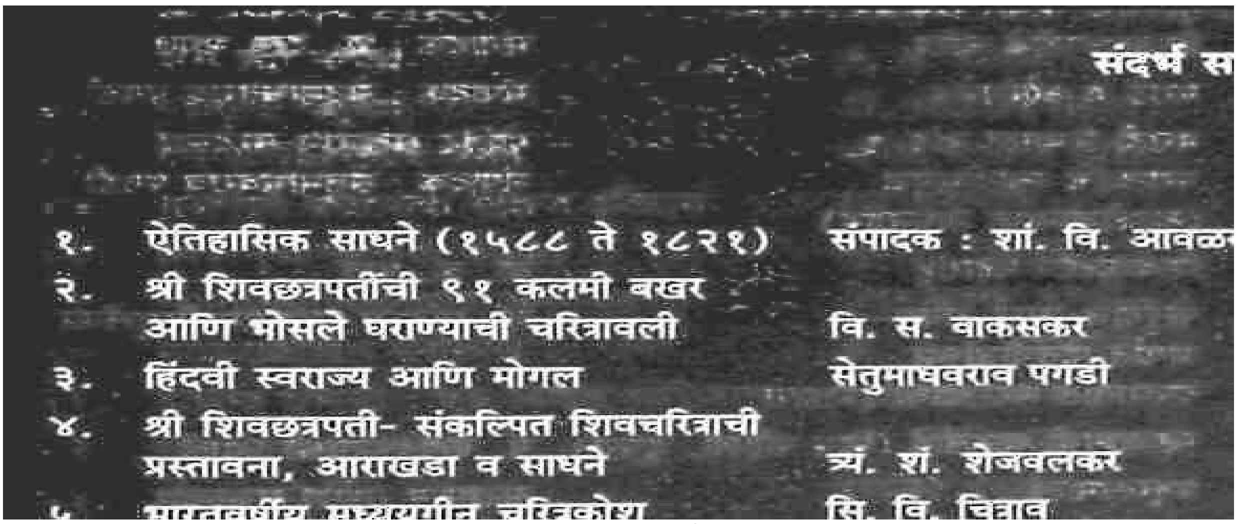
संदर्भ स
१. ऐतिहासिक साधने (१५८८ ते १८२१) संपादक : शां. वि. आवळ
२. श्री शिवछत्रपतींची ९१ कलमी बखर
आणि भोसले घराण्याची चरित्रावली वि. स. वाकसकर
३. हिंदवी स्वराज्य आणि मोगल सेतुमाधवराव पगडी
४. श्री शिवछत्रपती- संकलित शिवचरित्राची
प्रस्तावना, आराखडा व साधने त्र्यं. शं. शेजवलकर

Fig. 4 Final Output Of Project

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 1, January 2016

V. CONCLUSION

This paper presents an adaptive image contrast based document image binarization technique that is tolerant to different types of document degradation such as uneven illumination and document smear. The proposed technique is simple and robust, only few parameters are involved. Moreover, it works for different kinds of degraded document images. The proposed technique makes use of the local image contrast that is evaluated based on the local maximum and minimum. The proposed method has been tested on the various datasets. Experiments show that the proposed method outperforms most reported document binarization methods in term of the F-measure, pseudo F-measure, PSNR, NRM, MPM and DRD.

REFERENCES

1. B. Gatos, K. Ntirogiannis, and I. Pratikakis, "ICDAR 2009 document image binarization contest (DIBCO 2009)," in Proc. Int. Conf. Document Anal. Recognit., Jul. 2009, pp. 1375–1382.
2. Pratikakis, B. Gatos, and K. Ntirogiannis, "ICDAR 2011 document image binarization contest (DIBCO 2011)," in Proc. Int. Conf. Document Anal. Recognit., Sep. 2011, pp. 1506–1510.
3. Pratikakis, B. Gatos, and K. Ntirogiannis, "H-DIBCO 2010 handwritten document image binarization competition," in Proc. Int. Conf. Frontiers Handwrit. Recognit. Nov. 2010, pp. 727–732.
4. S. Lu, B. Su, and C. L. Tan, "Document image binarization using background estimation and stroke edges," Int. J. Document Anal. Recognit. vol. 13, no. 4, pp. 303–314, Dec. 2010.
5. B. Su, S. Lu, and C. L. Tan, "Binarization of historical handwritten document images using local maximum and minimum filter," in Proc. Int. Workshop Document Anal. Syst., Jun. 2010, pp. 159–166.
6. J. Bernsen, "Dynamic thresholding of gray-level images," in Proc. Int. Conf. Pattern Recognit., Oct. 1986, pp. 1251–1255.
7. J. Sauvola and M. Pietikainen, "Adaptive document image binarization," Pattern Recognit., vol. 33, no. 2, pp. 225–236, 2000.
8. W. Niblack, An Introduction to Digital Image Processing. Englewood Cliffs, NJ: Prentice-Hall, 1986.

BIOGRAPHY

UmeshSonawaneis BE, Student at Department of Information Technology,Raisoni College of Engineering,Ahmednagar, Maharashtra, India

Ajay Waghis BE, Student at Department of Information Technology,Raisoni College of Engineering,Ahmednagar, Maharashtra, India.

Pritam Bhandariis BE, Student at Department of Information Technology,Raisoni College of Engineering ,Ahmednagar, Maharashtra, India.

Harshada Puneis BE, Student at Department of Information TechnologyRaisoni College of Engineering,Ahmednagar, Maharashtra, India.

Pratik Kalamkar Assistant Professor at, Department of Information Technology Engineering,Raisoni college of Engineering Pune, Maharashtra, India.